Tecnicatura Universitaria en Inteligencia Artificial

Procesamiento del Lenguaje Natural(NLP)

Informe Trabajo Práctico Final

Alumna: Sol Kidonakis

Informe

Ejercicio 1

El objetivo principal de este proyecto es desarrollar un sistema de procesamiento del lenguaje natural (PLN) que extraiga, limpie y procese textos de documentos PDF y archivos CSV, generando embeddings para posteriormente almacenarlos en una base de datos ChromaDB. Adicionalmente, se desarrolló un chatbot experto utilizando la técnica RAG (Retrieval-Augmented Generation) para responder a consultas basadas en información de los documentos y una base de datos de grafos.

Métodos y Herramientas Utilizadas

1. **Extracción y Limpieza de Textos:**
    - Se utilizaron bibliotecas como `fitz` para extraer texto de archivos PDF.
    - La limpieza del texto se realizó utilizando expresiones regulares y `nltk` para tokenización y eliminación de stopwords.
2. **División de Texto en Fragmentos:**
    - Se empleó `Langchain` para dividir el texto en fragmentos de tamaño específico, facilitando el manejo y procesamiento de grandes volúmenes de texto.
3. **Generación de Embeddings:**
    - Se utilizó el modelo preentrenado `Word2Vec` de Gensim para generar embeddings promedio de los fragmentos de texto y datos del archivo CSV.
4. **Almacenamiento en ChromaDB:**
    - Los embeddings generados se almacenaron en una colección de ChromaDB para facilitar su recuperación y uso en el chatbot.
5. **Clasificación y Evaluación:**
    - Se desarrollaron dos clasificadores: uno basado en LLM utilizando palabras clave y otro entrenado con embeddings y un modelo de regresión logística.
    - Se evaluó el rendimiento del clasificador basado en regresión logística utilizando métricas de precisión, recall y matriz de confusión.

6. **Chatbot Experto con RAG:**
    - Se implementó un chatbot que utiliza técnicas de recuperación de información en ChromaDB, archivos CSV y una base de datos de grafos para responder a consultas.
    - La generación de respuestas se realizó utilizando el modelo `GPT-2` de Hugging Face.

Resultados

- **Extracción y Limpieza de Textos:** Se extrajeron y limpiaron con éxito los textos de los archivos PDF `Argentina.pdf` y `Brasil.pdf`, así como los datos del archivo CSV `WorldPopulation2023.csv`.
- **Generación de Embeddings:** Se generaron embeddings precisos para los fragmentos de texto y los datos del CSV, almacenándolos efectivamente en ChromaDB.
- **Clasificación:** El clasificador basado en regresión logística mostró un rendimiento aceptable con una precisión significativa en la clasificación de consultas.
- **Chatbot:** En este caso,las consultas no me funcionaron como deseaba,y por cuestiones de tiempo y entrega no me termino de funcionar.

Ejercicio 2

Rerank es una técnica para reordenar los resultados obtenidos de una búsqueda inicial o recuperación de documentos para mejorar la relevancia final. Primero se recuperan resultados preliminares basados en un modelo de recuperación (ej. búsqueda por palabra clave) y luego se aplican técnicas de reranking (ej. modelos de clasificación o regresión) para ajustar el orden de los resultados según un criterio más refinado.

Impacto en el Desempeño:
La técnica de reranking mejora la precisión al asegurar que los resultados más relevantes sean priorizados, lo cual puede llevar a respuestas más precisas y útiles para el usuario.

Aplicación:
Implementaría reranking en la fase donde los resultados se recuperan de las bases de datos (tanto tabulares como de grafos). Después de obtener los documentos relevantes, aplicar un modelo de reranking para ajustar la relevancia final antes de generar la respuesta del chatbot.

Diagrama: